Unsupervised Prosodic Break Detection in Mandarin Speech

Jui-Ting Huang[1], *Mark Hasegawa-Johnson*[1], *Chilin Shih*[2]

[1]Department of Electrical and Computer Engineering; [2]Departments of EALC/Linguistics
University of Illinois at Urbana-Champaign, USA

{jhuang29; jhasegaw; cls}@uiuc.edu

Abstract

We propose that, in Mandarin speech, an automatic prosodic break detector can be trained without any prosodically labeled training data. We use only lexical and acoustic cues to create a small labeled training set, then use semi-supervised learning to train a prosodic break detector. A generative mixture model is proposed as the learning algorithm that can learn with both labeled and unlabeled data. The experiments in both English and Mandarin corpus verify our algorithm.

1. Introduction

Prosodic breaks are boundaries which mark the perceived degree of separation between a pair of lexical items in human speech. A prosodic break detector is actually a classifier which receives acoustic correlates and classifies the event as *non-break* or *break*. Traditionally automatic prosodic labeling is based on supervised training methodology, in which data marked with prosodic events are required to train a classifier [1, 2]. All of them are supervised classification tasks that try to map acoustic or/and lexical cues to the prosodic event defined in TOBI [3] or any other system of prosodic annotation, given a well-labeled database such as [4].

The goal of this research is to automatically locate prosodic breaks in Mandarin speech *without any prosodically labeled data*. The advantage of our proposal is the prosodic structure can be detected for any Mandarin corpus regardless of the existence of prosodic labels. Furthermore, the prosodic structure detected is totally driven by the distribution of acoustic features. This provides an interesting view of how non-expert people perceive prosody without the labeling instruction, and how this *natural* prosodic structure interacts with acoustic and phonetic structure, as we human seem to process prosody information without a guideline being taught, too. This work can also aid future research in speech recognition and understanding.

There is little research about learning prosodic events unsupervisedly. Levow [5] employed both unsupervised and semi-supervised (with the aid of a small set of labeled data) in English pitch and Chinese tone recognition, using only acoustic features. Ananthakrishnan et al. [6] used clustering algorithm to partition the acoustic space into two contrastive classes and used lexical and syntactic to further refine the classification. In their approach, some reliable representatives of each cluster are identified by assuming the acoustic confusion associated with reliable samples is small, and use these representatives to train a classifier. We use the similar two-step strategy in this work, but there are two main differences between our work and [6]. First, we identify some representatives by not only acoustic but also lexical cues, and the reliability of those representatives is justified by the true characteristic in Mandarin speech rather than the plausible assumption about the acoustic confusion. Second, We train a classifier from both the identified labeled set and a large pool of unlabeled set, expecting the data structure embedded in the unlabeled set can further provide more information. Learning with both labeled and unlabeled data is called semi-supervised learning, and people have found the possibilities of the aid of unlabeled data in learning problems.

2. Proposed Method

2.1. Creating the labeled set

Based on the experimental study in the literature [7, 8], we adopt one key characteristic of fluent Chinese utterances: there is no prosodic break at a syllable boundary within a short lexical word. Here we define "short" as containing less than three syllables. By collecting these "intra-short-word" syllable boundaries, we have the set of data examples from the *non-break* class. The task now is find the *break* class among the rest of "inter-word" and "intra-long-word" syllable boundaries, given the *non-break* set that we have collected. Formally, we have a labeled set $\mathcal{X}_l$ and an unlabeled set $\mathcal{X}_u$:

Dataset I

$\mathcal{X}_l$ all "intra-short-word" syllable boundaries, with the corresponding class labels $y_i = nb$.

$\mathcal{X}_u$ all "inter-word" and "intra-long-word" syllable boundaries, where class labels $y_i \in \{nb, b\}$ are missing.

Unambiguous examples of the break class are also available, but only in certain speaking styles. In speech with few disfluencies (e.g., radio announcer speech), almost every silent pause is a prosodic phrase boundary. In speech of this style, we can say that a silent pause is a sufficient condition for the presence of a prosodic break at the corresponding syllable boundary. If we apply this rule beforehand, then a certain amount of break data can be obtained. Together with the non-break data extracted from the intra-short-word boundaries, the labeled set now have data from both classes, while we still have an unlabeled set which comprises inter-word syllable boundaries without silent pause. Different from Dataset I, the labeled set now have two subsets from each class:

Dataset II

$\mathcal{X}_{l,nb}$ all "intra-short-word" syllable boundaries, with the corresponding labels $y_i = nb$.

$\mathcal{X}_{l,b}$ all syllable boundaries that have silent pauses, with the corresponding labels $y_i = b$.

$\mathcal{X}_u$ The rest of "inter-word" and "intra-long-word" syllable boundaries, where class labels $y_i \in \{nb, b\}$ are missing.

2.2. Learning with both labeled and unlabeled data

The generative mixture model is proposed here to model both non-break and break data in the feature space, and it applies

to both Dataset I and II proposed in section 2.1. The model has M mixture components, $\mathcal{M}_1, \ldots, \mathcal{M}_M$, that can generate data (x, c, L) where $x \in \mathbb{R}^n$ is prosodic feature containing n acoustic correlates, $c \in \{nb, b\}$ is the class label, and $L \in \{\text{“}l''\text{, “}m''\}$ indicates label is observed (“l”) or missing (“m”). The joint probability of (x, c, L) is a weighted sum over all mixtures:

$$P(x, c, L) = \sum_{j=1}^{M} \alpha_j f(x|\theta_j) P(C = c|\mathcal{M}_j) P(L|C = c) \quad (1)$$

where α_j is the weight of mixture $\mathcal{M}_j$, $f(x|\theta_j)$ describes the probability of the feature x in mixture $\mathcal{M}_j$, (Here we use Gaussian distribution.) $P(C = c|\mathcal{M}_j)$ is the class probability in mixture $\mathcal{M}_j$, and $P(L|C = c)$ is the class-dependent label present/absent probability.

Therefore, the parameter set that we need to estimate now is $\Theta = \{\{\alpha_k\}, \{\theta_k\}, \{P(C = c|\mathcal{M}_j)\}, \{P(L = v|C = c)\}\}$, and we use Expectation-Maximization to find the optimal parameter that can maximize the data loglikelihood. It is a iterative algorithm where each iteration has two steps:

E-step estimate the expectation of complete-data log-likelihood with respect to the missing values

$$Q(\Theta, \Theta^{i-1}) = E_{\mathcal{X}_m}\left[\log p\left(\mathcal{X}, \mathcal{X}_m|\Theta\right)|\mathcal{X}, \Theta^{i-1}\right] \quad (2)$$

M-step update the parameter with the value which can maximize the auxiliary function in E-step:

$$\hat{\Theta} = \arg\max_{\Theta} Q(\Theta, \Theta^{i-1}) \quad (3)$$

The closed form EM updating formulas are put in Appendix.

Once the model parameters are learned after EM iterations are done, the classification of a test (new) feature x is by choosing the class label that maximizes the posterior probability:

$$\hat{C} = \arg\max_{c \in \{nb, b\}} P(c|x), \quad (4)$$

where the posterior probabilities can be obtained as

$$P(C = c|x) = \frac{\sum_{l=1}^{M} \alpha_l p(x|\theta_l) p(C = c|M_l)}{\sum_{l=1}^{M} \alpha_l p(x|\theta_l)}. \quad (5)$$

Also, the overall class distribution can be estimated as

$$P(C = c) = \sum_{l=1}^{M} \alpha_l p(C = c|M_l) \quad (6)$$

3. Experimental Results

3.1. Pilot Study with English Corpus

In addition to Mandarin, we use English prosody corpus as a pilot study of our algorithms. We use a subset of the Boston Radio News Corpus [4], read by female speaker F2B, comprising 34 stories, 49 minutes of news material. The corpus includes orthographic transcription and automatically generated phone alignments. The prosodic structure is annotated with the perceptual labeling system developed by Price et al. in [9]. Under this system, the degree of perceived disjuncture between each pair of words is expressed by a break index between 0 and 6, assigned by human labeler. The description for each break level are listed in Table 1. Because our classifier deals with *non-break* versus *break*, we merge 0 and 1 into the *non-break* set, and 2 to 6 into the *break* set. The acoustic features includes pitch, duration, energy related features, totally 99 variables. The feature dimension is reduced to two by Principal Component Analysis (PCA). We separate the last six stories out as the testing set. We use the English prosody data as the toy data to test the performance of the generative mixture model when the labeled set has only one class of samples (Dataset I) or both classes (Dataset II). To simulate Dataset I, we take a fraction $R = 0.5$ of non-breaks as a labeled set $\mathcal{X}_l$ and combine the rest of non-breaks with all breaks to form the unlabeled set $\mathcal{X}_u$. As for Labeled set II, in addition to a non-break labeled set $\mathcal{X}_{l,nb}$ is collected as in Dataset I, we take some break data to form $\mathcal{X}_{l,b}$, and the rest become $\mathcal{X}_u$.

Table 1: *Description for break indices by Price et al. [9].*

Break Index	Description
0	between two orthographic words between which there is obvious phonetic reduction
1	a default to unmarked word boundaries
2	a perceived grouping of words that is not intonationally marked
3	intermediate phrase boundaries
4	intonational phrase boundaries
5	perceived groupings of intonational phrases within a (typically long) sentence
6	sentence boundaries

For Dataset I, the generative model gives the overall non-break versus break classification accuracy 53.26%, which is near the chance rate 51.79%, but it is rather lower than the supervised result (70.77%) where the labels are given for training. The correct classification rates for each break level are listed in Table 2. As we can see from the table, the recognition of breaks is increasing as the perception strength of break is larger, and the sentence boundaries are recognized best among them. The indistinguishableness of break index 2 and 3 hurt the overall accuracy. In addition, the class distribution estimated using (14) is $0.5973 : 0.4027$, while the true sample class distribution is $0.562 : 0.438$.

For Dataset II, we test with both balanced and imbalanced cases. Balanced case is where the class distribution in the labeled set equals the unlabeled set; Imbalanced case is not. The recognition accuracy for balanced case is 70.96% and imbalanced is 70.50%, which shows our mixture model is robust to the varied data distribution across labeled and unlabeled set. This is attributed to the label missing probabilities introduced in our generative model (1). Either the overall acuracy or the respective classification rates for each break level, the model learned from Dataset II has better performance than Dataset I. It is predictable because the labeled set now has data from the other class, providing more guidance of the class information.

3.2. Mandarin corpus

The Mandarin database is the corpus used for duration study for Text-to-Speech in Bell Laboratories [10]. This corpus contains 427 sentences from news material, recorded by a male Mandarin speaker from Beijing. The prosodic annotation includes prosodic word boundaries, minor breaks and major breaks, and we only focus on minor and major breaks here. For each syllabic boundary, a total of 25 acoustic features includes pitch, duration, and energy related features are extracted. Instead

Table 2: *English experiment for Dataset I (DSI) and Dataset II (DSII): the respective recognition accuracies (%) for each break level.*

	non-break		break					total
Break index	B0	B1	B2	B3	B4	B5	B6	
DSI	66.67	80.55	17.98	16.78	20.59	48.65	74.42	53.26
DSII	88.87	85.56	30.7	51	71.18	100	97.67	70.96

Table 3: *The nonbreak/break classification accuracies (%) of the mixture model, using Dataset I (DSI) and Dataset II (DSII) in the Mandarin testing set.*

	nonbreak	break w/o sp	minor break w/ sp	major break w/ sp	total
DSI	70.32 (52.24)	83.03	86.67	87.19	73.25 (65.57)
DSII	83.98 (71.27)	66.07	76.87	90.8	81.93 (72.68)

of PCA, we use minimum Redundancy Maximum Relevance (mRMR) feature selection module [11] to select the top four variables as the input features. 52 sentences are separated out as the testing set.

Dataset I and II are built as described in section 2.1. Unlike English experiment, here we design a two-pass classification. Because major breaks and part of minor breaks are strongly correlated to silent pauses in this corpus, we treat the silent pause as a special feature and separate it from the feature set in the first pass. In the first pass, our algorithms detect breaks with all features excluding silent pause; in the second pass, we classify all data points with silent pauses into the *break* class. The reason of the two-pass classification is to investigate how semi-supervised learning will learn the prosodic structure with acoustic cues other than silent pause.

For Dataset I, the overall classification accuracy after the two-pass classifier is 75.76%, whereas the same two-pass classifier but with a supervised classification where prosodic labels are given for training, in the first pass, will give 88.62%. Because the generative model is applied to the first pass, we are also interested in the classification result in the first pass, and the accuracy for each break level in the first pass is listed in Table 3. Breaks are redivided into three categories–minor breaks without silent pauses, minor breaks with silent pauses, and major breaks with silent pauses. The algorithm recognizes those breaks quite well even without given the silent pause cue. The estimated class distribution is $0.51 : 0.49$ as the real distribution is $0.7869 : 0.2131$.

For Dataset II, the overall accuracy (85.50%) is higher than Dataset I case, and comparable to the supervisd result, 88.62%. The accuracy for each break level in the first pass is also listed in Table 3. More percentages of nonbreaks are detected than Dataset I, and the classification rates varies across different type of breaks unlike Dataset I in which the rates are very similar. The estimated class distribution is $0.6411 : 0.3589$, which is closer to the real one than Dataset I.

4. Discussion

For Dataset I, Mandarin case seems to have much higher recognition of breaks than English. One interpretation is that Mandarin break and non-break may have simpler data structure to capture. Both English and Mandarin prosody data show that Dataset II improves over Dataset I in overall recognition accuracy. When we look in detail at Table 3, there are actually some differences in how different types of breaks are classified. In English data, the addition of some breaks in the labeled set, i.e., Dataset II, improves the classification rates for all levels of breaks and nonbreak. In Mandarin data, only the nonbreak and major break with silent pause are improved where as the other kinds are degrading. The possible reason is that in Mandarin we use the breaks with silent pauses as the seeds for break class to run EM, resulting in a high recognition rate for the same type of break and also nonbreak. Nevertheless, from both experiments we can clearly see that having a labeled set from both classes can learn a better model than just from one of the classes.

Mandarin is a tonal language; every syllable has a tone signaled by different pitch contour shapes. In this paper we haven't considered the influence of tone into the prosodic break detector. However, the detector still works because the most relevant prosodic features selected automatically by mRMR algorithm are not about the shape of the pitch contour, which are the most relevant features to tone differentiation.

It is hard to compare English and Mandarin results fairly; English prosody data are used to detect break among word boundaries while Mandarin prosody data are used to detect break among syllable boundaries. To provide a fairer comparison, we also calculate the recognition accuracy concerning only word boundaries. That is, the denominator of accuracy does not include "intra-word" syllable boundaries. The resulting different numbers are parenthesized in Table 3. The accuracy is worse, which implies that the nonbreaks that occur between two words (inter-word) are worse recognized than those within a word (intra-word).

Different measures of the classification results will provide different information. The rate of total accuracy does not provide enough information for us to understand what the algorithm has learned, and that is why we also look into the classification rate for each type of breaks for more details. It is possible that other evaluation metrics might also help to analyze classifier performance, such as F-measure or break insertion and deletion rate.

5. Conclusion

We propose that in Mandarin speech prosodic breaks can be located without any prosodically labeled data. The first method (Dataset I) makes use of non-break data which are obtained from "intra-short-word" syllable boundaries, and learns its contrastive class with a generative mixture model. The second method (Dataset II) augments the non-break set with some break data that can be determined by silent pauses, and learns the mixture model given this set together with other unlabeled data. The generative mixture model is proposed to consider the

existence of both labeled and unlabeled data, and the prosodic break detector is based on the MAP rule using the generative model. When provided with only some labeled nonbreak data (Dataset I), the detector is able to discover the breaks from unlabeled data. In the experiments, Mandarin data has higher rate of discovering breaks than English. When provided with also some labeled break data (Dataset II), in both corpora it achieves a comparable recognition accuracy to the supervised case where all prosodic labels are given for training.

Since our approach learns classes only depending on the distribution of data in the corpora, it can be applied any corpus and automatically fit to the speaker-dependent or corpus-dependent spoken style, e.g. broadcast news, read speech, or telephone speech.

6. Acknowledgment

The authors would like to thank Dustin Hillard and Mari Ostendorf from University of Washington for the English data. This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-0534133 to Chilin Shih and Gary Cziko.

7. References

[1] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech and Language*, vol. 10, pp. 155–185, 1996.

[2] K. Chen, M. Hasegawa-Johnson, and A. Cohen, "An automatic prosody labeling system using ann-based syntactic-prosodic model and gmm-based acoustic-prosodic model," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, 2004, pp. 509–512.

[3] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "Tobi: a standard for labeling english prosody," in *Proceedings of the 2nd International Conference of Spoken Language Processing*, 1992, banff, Canada.

[4] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," Boston University, Tech. Rep. ECS-95-001, 1995.

[5] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *HLT-NAACL*, 2006, pp. 224–231.

[6] S. Ananthakrishnan and S. Narayanan, "Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling," pp. 297–300, 2006.

[7] M. Chu and Y. Qian, "Locating boundaries for prosodic constituents in unrestricted mandarin texts," *Computational Linguistics and Chinese Language Processing*, vol. 6, no. 1, pp. 61–82, 2001.

[8] Q. yao and P. wuyun, "Prosodic word: the lowest constituent in the mandarin prosody processing," pp. 591–594, 2002.

[9] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," in *HLT*, 1991.

[10] C. Shih and B. Ao, *Duration Study for the Bell Laboratories Mandarin Text-to-Speech System*. New York: Springer-Verlag, 1996, pp. 382–399.

[11] F. Long and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

A. EM updating formulas

The E-step computes the following two equations

$$p(l|x \in X_{l,c}, \Theta^g) = \frac{\alpha_l^g p(x|\theta_l^g) P^g(C=c|M_l) P^g(L="l"|C=c)}{\sum_{k=1}^{M} \alpha_k^g p(x|\theta_k^g) P^g(C=c|M_k) P^g(L="l"|C=c)} \tag{7}$$

and $\sum_{l=1}^{M} p(l|x \in X_{L,c}, \Theta^g) = 1, \forall c \in \{nb, b\}$;

$$p(l, C=c|x \in X_u, \Theta^g) = \frac{\alpha_l^g p(x|\theta_l^g) P^g(C=c|M_l) P^g(L="m"|C=c)}{\sum_{k=1}^{M} \alpha_k^g p(x|\theta_k^g) P^g(C=c|M_k) P^g(L="m"|C=c)} \tag{8}$$

and $\sum_{l=1}^{M} \sum_{c \in \{nb,b\}} p(l, C=c|x \in X_U, \Theta^g) = 1$.

For The M-step, we update the parameters by the following equations:

$$\alpha_l = \frac{1}{L+U} w_l, \tag{9}$$

where L is the size of the labeled set, U is the size of the unlabeled set, and

$$w_l = \sum_c \sum_{x \in X_{l,c}} p(l|x \in X_{l,c}, \Theta^g) + \sum_{x \in X_u} \sum_c p(l, C=c|x, \Theta^g). \tag{10}$$

The parameters of $f(x|\theta_l)$ are mean μ_l and covariance Σ_l:

$$\mu_l = \frac{1}{w_l} \Big(\sum_c \sum_{x \in X_{l,c}} x \cdot p(l|x \in X_{l,c}, \Theta^g) + \sum_{x \in X_u} \sum_c x \cdot p(l, C=c|x, \Theta^g) \Big) \tag{11}$$

$$\Sigma_l = \frac{1}{w_l} \Big(\sum_c \sum_{x \in X_{l,c}} p(l|x \in X_{L,c}, \Theta^g)(x-\mu_l)(x-\mu_l)^T + \sum_{x \in X_u} \sum_c p(l, C=c|x, \Theta^g)(x-\mu_l)(x-\mu_l)^T \Big). \tag{12}$$

Also, $P(C|M_l)$ and $P(L="m"|C)$ are updated in the following way:

$$P(C=c|M_l) = \frac{1}{w_l} \Big(\sum_{x \in X_{L,c}} P(l|x \in X_{l,c}) + \sum_{x \in X_u} P(l, C=c|x \in X_U) \Big). \tag{13}$$

$$P(L="m"|C=c) = \frac{U_c}{L_c + U_c}, \tag{14}$$

For Dataset I, the class labels for the break class are always missing, which means the labeling missing probabilities for the break class, $P(L="m"|C=b)$, equals 1. Therefore, the EM update formulas are the same as,(7) to (8), (9) to (14), except that $X_{l,nb} = X_l$ and $X_{l,b} = \phi$.